

APPLICATION FOR
UNITED STATES LETTERS PATENT
SPECIFICATION

Inventor(s): Yoshinobu HOTTA, Katsuhito FUJIMOTO,
Satoshi NAOI and Misako SUWA

Title of the Invention: CHARACTER STRING RECOGNITION
APPARATUS, CHARACTER STRING
RECOGNIZING METHOD, AND STORAGE
MEDIUM THEREFOR

CHARACTER STRING RECOGNITION APPARATUS, CHARACTER
STRING RECOGNIZING METHOD, AND STORAGE MEDIUM
THEREFOR

5 **Background of the Invention**

Field of the Invention

 The present invention relates to a character
string recognition apparatus in a character
recognition apparatus associated with a character
10 recognition apparatus OCR and software OCR which
have recently been more and more demanded as
document input appliances.

Description of the Related Art

15 As a method of recognizing a word by dividing
a handwritten character string into partial
character strings, there has been a suggestion for
recognizing a word in an area enclosed by the
characters (key characters) detected as delimiters
20 such as 都 (capital), 道, 府, and 県 (prefecture), 市
(city), 区 (ward), 郡 (county), etc. in a handwritten
address (for example, Japanese Patent Publication
No.11-161740, Japanese Patent Publication No.11-
328315).

25 However, the above mentioned method is used

for a handwritten address, and no method has been developed for a common handwritten character string other than an address. In an address, each key character can be a delimiter of character strings, and no plural characters have been used as key characters, that is, a key word.

In the conventional key character extraction system for use in recognizing a handwritten address, a key character is limited to {都 (capital), 道 (prefecture), 府 (prefecture), 県 (prefecture), 市 (city), 区 (ward), 郡 (county), 町 (town), 村 (village)}. However, when a target character string is not an address, it is necessary to newly set a key character each time a character string recognizing process is performed.

Furthermore, since the number of key characters is one (1) to be extracted from a character string in the conventional method, a key word containing two or more characters cannot be successfully extracted due to, for example, touching characters in the key word, if the conventional method is applied as is.

Additionally since a high precision rejecting process has been not performed in the conventional word recognizing method, a word can be

misrecognized as a completely wrong word, thereby discouraging users.

Summary of the Invention

5 The present invention has been developed to solve the problem with the above mentioned conventional technology, and aims at providing a character string recognition apparatus capable of automatically extracting a key word from a
10 character string, extracting a key word with high precision although the key word includes a plurality of characters, and preventing a word to be recognized from being misrecognized as quite a different word.

15 The character string recognition apparatus according to the present invention includes a recognition target character string group storage unit for storing a list of character strings in a category to be recognized; and a key word
20 determination unit for searching the recognition target character string group storage unit for each character to obtain the number of occurrences of each character, defining the character having a large number of occurrences as a key character, and
25 defining the character string having a large number

of occurrences as a key word.

Since a key character and a key word can be automatically determined depending on the category to be recognized, an operator, etc. does not have
5 to sequentially set a key character and a key word.

Furthermore, the character string recognition apparatus according to the present invention further includes a key character/word storage unit for storing a determined key character or key word;
10 and a key character/word extraction unit for extracting a character string as a key word if a part of the character string in the key word is extracted when a key character or a key word stored in the key character/word storage unit is extracted
15 from a character string to be recognized.

Thus, a key word can be extracted although no characters in a key word can be recognized, for example, because characters touch each other, etc. when the key word is extracted from a character
20 string to be recognized.

Furthermore, the character string recognition apparatus according to the present invention also includes a word recognition unit for recognizing a word, and a verification unit for checking whether
25 or not a recognition result of the word recognition

unit is correct.

For example, the above mentioned verification unit verifies a recognition result based on the line density or the peripheral distribution.

Thus, high precision recognition result can be provided for a user by verifying a word recognition result and rejecting a word recognition result if the word is misrecognized.

10 **Brief Description of the Drawings**

FIG. 1 shows the outline of the present invention;

FIG. 2 shows outline of the entire configuration of the character string recognition apparatus according to an embodiment of the present invention;

FIG. 3 shows embodiments of the automatic key character code extracting process and the key character/key word extracting process;

20 FIGS. 4A, 4B and 4C show examples of a character group to be recognized, and extracted characters and character strings;

FIG. 5 is a flowchart of an individual character segmenting process and a character recognition and key word extracting process;

FIG. 6 shows the case in which a character segmented as an individual character is recognized as a word;

FIG. 7 shows a verifying process by line density/peripheral distribution;

FIG. 8 is an explanatory view for a verifying process by line density/peripheral distribution;

FIG. 9 shows a configuration of the hardware of an information processing device; and

FIG. 10 shows an example of a storage medium.

Description of the Preferred Embodiment

A handwritten character string to be processed according to the present invention can be various character strings such as an address, a name, the name of a university, the name of a bank, etc. written on a normal listing. In the present invention, a character is not recognized one by one from the start of a character string, but a character string is divided into a plurality of partial character strings, and words contained in each partial character string are holistically recognized. Thus, characters touching each other and a separated character, which have been the problems with handwritten character strings, can be

correctly processed. The character recognition apparatus according to the present invention can be applied not only to the above mentioned handwritten character recognition apparatus, but also to a
 5 general character recognition apparatus such as a printed character recognition apparatus, a character recognition apparatus for use in a portable information terminal, etc.

FIG. 1 shows the outline of the present
 10 invention.

To solve the above mentioned problems, as shown in FIG. 1, a key character code extraction unit 1 automatically extracts a key character (key word) from a character string group (code) to be
 15 recognized, and enters the key character according to the present invention. Thus, as shown in FIG. 1 (and described later in detail by referring to FIG. 4), in addition to the key characters in address recognition such as 県 (prefecture), 市 (city), 区 (ward), 町 (town), etc., a key word, for example, 信用組合 (credit union), 支店 (branch office), 農業協同組合 (agricultural cooperative union), 支所 (branch), etc. are entered. In addition, by extracting a set of associated character strings, a set of key words
 20 having high probability of appearing together such

as 信用組合 (credit union) and 支店 (branch office), 農業協同組合 (agricultural cooperative union) and 支所 (branch), etc. are extracted.

Then, a key word extraction unit 2 segments an individual character from a character string image for character recognition, and extracts with high precision the character string corresponding to the key word extracted and entered by the key character code extraction unit 1 in combination with the language process. Then, a word area segmented by a key word is extracted from a character string image, and a word recognition unit 3 recognizes a word. Furthermore, a verification unit 4 verifies a word recognition result, and outputs a final character string recognition result.

FIG. 2 shows the outline of the entire configuration of the character string recognition apparatus according to an embodiment of the present invention.

In FIG. 2, in an automatic key character code extracting process 11, a key character code and a key character string (word) code are automatically extracted from a character string group to be recognized (character code group) (both character code and character string code can be hereinafter

referred to as a character code). Assume that a character string group to be recognized, for example, an address, a name, the name of a university, the name of financial facilities, etc. is designated (for example, set by a user).

In the automatic key character code extracting process 11, a character code appearing for a target character string group is checked, and a character having a large number of occurrences, and a character and a character string having large numbers of occurrences in a character string unit are entered as key character codes.

Furthermore, the correlation between characters is checked. The correlation refers to, for example, the probability of an occurrence of a character code B when a character code A appears, the probability of a simultaneous occurrence of character codes D and E when a character code C appears, etc. Thus, a set of character codes appearing with correlation is also entered as a key character code.

In the above mentioned process, { 県 (prefecture), 市 (city), 区 (ward), 町 (town) } or { 県 (prefecture), 郡 (county), 町 (town) } or { 都 (capital), 区 (ward), 町 (town) } can be automatically

extracted as key characters from an address. From the name of financial facilities, {銀行 (bank), 支店 (branch office)} {信用組合 (credit union), 営業部 (sales department)} {農業協同組合 (agricultural cooperative union), 支所 (branch)} can be automatically extracted as key characters. For example, in an address, a character which is not a delimiter of an address description can be used as a key character. A character code extracted as described above can be entered as a character code.

In a key character/key word extracting process 12, an individual character is segmented from a handwritten character string (image data) as described above, each of the segmented character is recognized, and a key character and a key word corresponding to the key character code extracted by the automatic key character code extracting process 11 are extracted and entered. Since the number of types of key characters is small, the processing time of key characters is shorter than that of normal characters (for example, about 4,000 types of characters are compared), and key characters indicate higher recognition precision than normal characters.

In a word area extracting process 13, a word

area delimited by the key character and key word is extracted from a handwritten character string (image). For example, if a handwritten character string is '東京信用組合 千駄木支店' (Tokyo Credit Union Sendagi Branch office), the areas of '東京' (Tokyo) and '千駄木' (Sendagi) delimited by '信用組合' (Credit Union) and '支店' (Branch office) are extracted as a word area.

In a word recognizing process 14, the character image of the extracted word area is normalized and a characteristic is extracted, and the characteristic is compared with the characteristic data of a word feature dictionary, etc. for word recognition. A word recognizing method can be, for example, methods disclosed by the above mentioned Japanese Patent Publication No.11-161740 and Japanese Patent Publication No.11-328315, and a method disclosed by Japanese Patent Application 2000-304758 previously suggested by the Applicant of the present invention.

As described above, a word is recognized for a word area enclosed by key words, and a word recognition result is verified as follows.

(i) In a character segmenting/character recognizing process 15, an individual character is segmented,

and a word recognizing process is performed. If a rate higher than a threshold is contained in the characters in a recognized word within n highest order of character recognition results, then a word recognition result by a word recognizing unit 14 is output. Otherwise, the result is rejected. For example, if '神奈川' (Kanagawa) is used as a word example, each of the characters '神', '奈', and '川' is recognized, n higher order probable recognition results are obtained for each character. If each character in a word as a word recognition result is not contained in the probable recognition results at or higher than a predetermined rate (for example, two or more characters in three characters), the result is rejected. Thus, apparent misrecognition can be reduced in the word recognition.

The value of n above can be set by, for example, an operator, etc. in advance.

(ii) Furthermore, as indicated by the dotted lines shown in FIG. 2, a word recognition result can be verified by comparing the numbers of characters in a character number estimating process 17.

That is, when word areas are holistically recognized, the number of characters in a recognized word is compared with the number of

characters estimated from a word image. If the difference between the numbers is equal to or larger than a threshold, then a word recognition result is rejected. As a method of estimating the number of characters, for example, a number computed based on the ratio of the height to the width of a bounding rectangular of the word image (the ratio of the height h to the width w in an example shown in FIG.2 = w/h). Otherwise, for example, peripheral distribution (black pixel projection histogram) is obtained for a word image, and a number obtained from the position of a delimiter of the peripheral distribution can be used.

(iii) In a verifying process 16 by line density/peripheral distribution, if a word feature is generated by combining character features, then the combination position of character features reciprocally computed as described later, a word image is divided in each position, and the line density and the peripheral distribution computed for each of the divided areas are computed with the predetermined line density and the peripheral distribution of each character of a recognized word.

If they are different, the word recognition result

is rejected, thereby verifying the word recognition result. As it is well known, the line density refers to the number of lines measured when a target character is scanned in the vertical or horizontal direction (Y or X direction).

Described below are the automatic key character code extracting (determining) process, the key character/key word extracting process, and the verifying process by line density/peripheral distribution.

(1) Automatic key character code extracting process, and Key character/key word extracting process

FIG. 3 shows an embodiment of the automatic key character code extracting process and the key character/key word extracting process.

First, in process 11a in the automatic key character code extracting process 11, a character having a large number of occurrences, a character and a character string having large numbers of occurrences in a character string unit are extracted as a key character code from a code of a character string to be recognized. Furthermore, in process 11b, a set of character codes appearing with correlation is extracted, and a set of

characters is entered as a key character code.

The above mentioned 'character having a large number of occurrences in a character string unit' can be, for example, as shown in FIG. 4A, '県 (prefecture)' because it appears in a number of character strings in '愛知県海部郡美和町3-3-3' (3-3-3, Miwa-town, Kaifu-county, Aichi-prefecture) '長野県信用組合松本支店' (Nagano Credit union Matsumoto Branch office), etc. The 県 (prefecture) frequently appears in the entire character string group 20 to be recognized. Thus, there can be a character frequently appearing in both groups.

FIGS. 4A through 4C show examples of a character string group to be recognized, and key character codes extracted therefrom.

FIG. 4A shows an example of a list 20 of a character string group to be recognized. FIG. 4A shows a character string, but a code of each character in the character string group is actually used.

The list of the character string group 20 to be recognized contains character strings as recognition results in the category to be recognized (for example, an address, a name, the name of a university, the name of financial

facilities, etc. as described above).

In the example shown in FIG. 4A, for example, a client/banker, etc. writes the name of financial facilities (name of a bank and the name of its branch office) and its address in an actual use on a predetermined form, and a character recognition apparatus reads/recognizes them. In this use, the names of actual financial facilities to be recognized (name of a bank and the name of its branch office) are listed. That is, the client, etc. enters on a form one of the lists shown in FIG. 4A, for example, '東京信用組合千駄木支店 東京都文京区千駄木1-1-1'(Tokyo Credit union Sendagi Branch office 1-1-1, Sendagi, Bunkyo-ward, Tokyo-Capital) , etc. Therefore, these lists can be generated in advance, based on which the present embodiment performs the following process.

First, using the above mentioned list of the character string group to be recognized, the number of occurrences of each character (character code) is checked in the list. Then, a character code having a large number of occurrences (for example, up to m highest order codes; the value of m is determined set in advance by an operator, etc.) is extracted as a key character code. However, a key

character is extracted not only as one character, but also as two characters (hereinafter possibly referred to as a key character string). When characters having large numbers of occurrences continuously appear, they are extracted as a key character string. For example, in a list shown in FIG. 4A, characters having large numbers of occurrences such as '県', '市', '支', '店', '所', etc. are extracted, '県' and '市' have no other characters having large numbers of occurrences and continuously appearing before or after them. Therefore, each of them is individually extracted as a key character. On the other hand, since '支', '店', and '所' continuously appear as '支店', '支所', etc., it is extracted as a key character string. Similarly, for example, '信用組合', '農業組合', etc. are extracted as key character strings.

For example, assuming that a character code having a large number of occurrences is extracted from a character string group to be recognized shown in FIG. 4A, for example, as shown in FIG. 4B, key character codes such as the names of financial facilities, agricultural cooperative union, etc. such as '信用組合', '商工信用組合', etc. are extracted in addition to the characters in recognizing an

address expressed by 県, 都, etc. Furthermore, if the number of occurrences of a character string indicated with a character such as 県, etc. such as '... 県信用組合', '... 県農業協同組合', etc., is large, then it is also extracted as a key character code. When the correlation of appearances of characters is checked, a set of character codes indicating high correlation is extracted as shown in FIG. 4C.

The method of checking correlation (practical processing method) can be, for example, counting for each character code in a character string to be recognized, which contains the character, other character codes, accumulating the values, and extracting a character having the largest accumulation value or having a value larger than a predetermined threshold (set by an operator, etc. in advance). For example, when a key character '都' is used as an example, the first '都' appears in the character string to be recognized '東京都文京区千駄木1-1-1' in the list shown in FIG. 4A. Therefore, other character codes '東', '京', '区', '千', '駄', and '木' are counted in association with '都' (numbers/symbols such as 1-1-1, etc. are not counted). If this process is performed on all character strings to be recognized shown in FIG. 4A,

the count value (accumulation value) of '区' is the largest for '都'. Therefore, '都'-'区' is extracted as a set of character codes having high correlation. When this process is performed not only on each character code, but also on each character string, for example, '信用組合'-'支店', '農業組合'-'支所', etc. are extracted as a set of character strings having high correlation.

The above mentioned process can be performed on all characters and character strings, but it is more effective to perform the process on the already extracted key character or key character string.

When a character which is rarely misrecognized in character recognition is entered in advance and the key character code is extracted, a key word can be more correctly extracted in the process of extracting a key word from a character image.

In an individual character segmenting process 12a in the key character/key word extracting process 12, a character is individually segmented from a handwritten character string (image data), and each character is recognized as described above, and a key character and a key word corresponding to a key character code extracted by the automatic key

character code extracting process 11 are extracted.

Described below is the process of extracting a plurality of key characters, that is, a key word. For example, using the name of financial facilities, '信用組合' and '農業協同組合' are extracted as key words from '〇〇 信用組合△△支店', '〇〇農業協同', etc. A key character such as 県, 市, 区, 町, etc. can be extracted in the key character/key word extracting process described below as a key word is extracted.

The process of extracting a common key word is described below. In a character recognizing and key character/word extracting process 12b, a segmented individual character is recognized, and the distance values of n probable higher order recognition results are checked for each character. If a character in a key word entered in advance belongs to probable recognition results whose distance values are equal to or smaller than a threshold TH1, then it is a probable key character of a target character.

When there are a plurality of relevant characters in a key word, a plurality of probable key characters are presented. The process is performed on all characters. When the same character string as the key word is contained in a

sequence of probable key characters, the character string is extracted as a key word.

FIG. 5 is a flowchart of the above mentioned individual character segmenting and key character/key word extracting process.

In step S1, an individual character is segmented, and the segmented character is recognized in step S2. The character is recognized by extracting a characteristic of a segmented character, extracting a probable character by comparing the characteristic data with a stored dictionary, and obtaining a value of the distance between the segmented character and the probable character.

In step S3, the n probable higher order characters of the character recognition results obtained as described above are extracted. In step S4, it is checked whether or not the distance value of a probable character is smaller than the threshold TH1. If the distance value of the probable character is smaller than the threshold TH1, then it is checked in step S5 whether or not the probable character is contained in the key word entered in the automatic key character code extracting process 11. If yes, the character is

entered as a probable key character in step S6.

The above mentioned process is repeated until the process is performed on all characters. When the process is performed on all characters, control is passed from step S7 to step S8, and a sequence of probable key characters (a combination of characters) containing the same character string as the key word entered in the automatic key character code extracting process 11 is extracted as a key word.

For example, assume that a character string 'o Δ' (unknown characters) is contained in a character string image to be recognized, '銀' and '錢' are entered as probable key characters for the character 'o' , and '行' and '桁' are entered as probable key characters for the character 'Δ' in the process in steps S4 through S6. Assuming that only '銀行'(bank) is entered as a key word in the automatic key character code extracting process 11 in the combinations of the probable key characters ('錢桁', '銀桁', '錢行', and '銀行') ('錢' and '桁' are considered to be used in completely different key words entered), the character string 'o Δ' is recognized as the key word '銀行'.

On the other hand, when the above mentioned

character recognition is performed, a word recognizing process is performed on a character segmented as an individual character in a word recognizing/key word extracting process 12c.

5 For example, when small written characters touch each other as the '支店' (branch office) as shown in FIG. 6, the touching characters can be misrecognized as one character when characters are segmented.

10 In this case, in the word recognizing/key word extracting process 12c, a comparing process is performed on an area segmented as a character in word feature as well as character feature. When a distance condition is satisfied in the comparison
15 in word feature, the word is extracted as a key word.

 When a part of characters in a key character entered in advance are extracted in the above mentioned key character/key word extracting process,
20 the following process is performed.

(i) 2-step extraction using multiple stage threshold

 When only a part of a key word is extracted from a character string, a distance value condition
25 of the key word extraction is mitigated on the

preceding/subsequent character recognition results in a preceding/subsequent key character extracting process 12d, and the extracting process is performed again.

5 That is, although a probable key character is extracted when a key character is contained in probable recognition results having distance values equal to or larger than TH1 after obtaining character recognition results from a character, a

10 key character is extracted from probable recognition results having distance values equal to or larger than TH2 ($TH2 > TH1$) for the preceding/subsequent characters of an extracted key character. Thus, even if a part of characters of a

15 key word indicate considerable deformation, the characters can be extracted as a part of the key word.

20 Values that an operator, etc. considers appropriate are set as the values of TH1 and TH2 based on, for example, the past experiment results, etc.

(ii) Key character extracting process by recognition of both ends

25 In a character string containing a number of touching characters, it is often difficult to

segment and recognize all individual characters belonging to key words

For example, in a key word image '農業協同組合', the characters '業協' complicatedly touch each other, and cannot be successfully segmented in the character segmenting process. Then, a key character extracting process 12e by recognition of both ends is performed.

In the key character extracting process 12e, the leading and trailing characters of a character string are extracted for a key word containing N or more characters, and when P% or more characters of the entire characters are extracted, it is assumed that a key word has been extracted. The above mentioned N and P are the values obtained in the experiment. For example, P = 60.

In '農業協同組合', for example, when the probable key characters '農業oo組合' and '農oo同組合' are extracted, the character string is defined as '農業協同組合'. (In this example, it is assumed that no other similar words ('農林協同組合', etc.) exist.)

(iii) Key word extracting process by partial recognition

When the leading character of a key word touches the preceding character, the key word

cannot be correctly extracted, and the above mentioned both end recognition system cannot be successfully applied. Therefore, a key word extracting process 12f by partial recognition is performed.

In the key word extracting process 12f by partial recognition, when a part of characters of the first and second half of a character string are recognized for key words each containing M or more characters, the key character extracting process 12e by recognition of both ends ((ii) above) is performed on both ends of partial character string, and the partial character string is extracted as a partial key word when the condition is satisfied (when P% or more characters are extracted).

In '農業協同組合', for example, when 'ooo業△△組合' is recognized, the key word contained in the character string is estimated to be '農業協同組合'. However, since the position of the '農' is uncertain, (that is, in this case, since there is the possibility that the '農' touches any preceding character, it cannot be clearly recognized which part of the touching characters belongs to '農'), the portion subsequent to the '業' is determined as '業協同組合'.

As described above, if a key word is extracted as described above, the extracted key word is verified in a key word verifying process 12g by word recognition.

5 In the key word verifying process 12g by word recognition, a word feature of a key word is generated when the key word is extracted. Then, a word recognizing process is performed on a key word area extracted in the key word extracting process, and only the areas satisfying the distance condition are extracted as key words.

Furthermore, some word images can be easily misrecognized in comparing word features in combinations of characters other than key words.

15 Then, these easily misrecognized word images are added to a word feature dictionary as similar word features for use in identifying from correct word features, thereby improving the extraction precision of the key word.

20 (2) Verifying process by line density/peripheral distribution

As described above by referring to FIG. 2, a word area delimited by a key word is extracted, a word is recognized in the word area delimited by the key word, character segmentation and

verification by character recognition are performed, a word recognition result is verified by assuming the number of characters, and the verifying process by line density/peripheral distribution is performed on a word recognition result.

The verifying process by line density/peripheral distribution is described below by referring to FIGS. 7 and 8. In this example, in the word recognizing process in the word recognizing process 14, a word is recognized by generating a word feature by a combination of character features, and comparing the word feature with the characteristic of the word image in the extracted word area.

Since the 'line density' and the 'peripheral distribution' are well known, the detailed explanation is omitted here.

When a word recognition result is obtained by the word recognizing process 14, the combination position of character features is reciprocally computed in a character feature combination position computing process 16a. That is, the combination position of character features is reciprocally computed from a word template used when a word is compared in the system of

recognizing a word by generating word features.

For example, as shown in FIG. 8, when '七十七' (seventy-seven) is obtained as a comparison result of a word image '富士'(Fuji), the combination position is reciprocally computed from the word template, and the combination position of each character of '七十七' is obtained.

Since the combination position is the position in the word normalized image, the position is converted into the position in a word image, and a word area is divided in the word image in each of the converted positions in a word area dividing process 16b. For example, as shown in FIG. 8, the combination position of each character of '七十七' is converted into the position on the word image, and the word image '富士' is divided.

In a line density and peripheral distribution computing process 16c, the line density or the peripheral distribution is computed for each of the divided areas. For example, in the example shown in FIG. 8, the line density of each of the divided areas of the word image '富士' is computed.

On the other hand, in a line density and peripheral distribution computing process 16d, the line density or the peripheral distribution of each

character is computed for each character of a word recognition result by referring to the line density dictionary 31, the peripheral distribution dictionary 32, etc. For example, in the example shown in FIG. 8, the line density of each character of the word '七十七' is extracted.

Then, in a comparing process 16e, the line density or the peripheral distribution computed for each of the divided areas is compared with the line density or the peripheral distribution of each character of the recognized word. If they do not match, the word recognition result is rejected.

The difference level for rejection (threshold) is set by an operator, etc. in advance based on an experiment result, etc.

In the example shown in FIG. 8, since the line density in each area obtained by dividing a word image '富士' is apparently different from the line density of each character of the word '七十七', the word recognition result '七十七' is rejected.

FIG. 9 shows an example of a hardware configuration of the information processing device realizing the character string recognition apparatus according to an embodiment of the present invention.

The character string recognition apparatus according to an embodiment of the present invention can be realized by a normal computer system as described above. (The apparatus is not limited to this system. That is, it can be realized by a portable information processing device, etc. such as a PDA/hand-held PC, etc.)

An information processing device 40 shown in FIG. 9 comprises a CPU 41, memory 42, an input device 43, an output device 44, an external storage device 45, a medium drive device 46, a network connection device 47, etc. interconnected through a bus 48. Furthermore, an optical reader 50, etc. is connected to the bus 48. The configuration shown in FIG. 9 is an example, and the device is not limited to this configuration.

The optical reader 50 is, for example, a scanner, a digital camera, etc., and transfers digital image data obtained by optically reading a character, etc. on paper to the information processing device 40. (This refers to the above mentioned 'character string image' and 'handwritten character string (image)').

The CPU 41 is a central processing unit for controlling the entire information processing

device 40.

The memory 42 can be RAM, etc. for temporarily storing a program or data stored in the external storage device 45 (or a portable storage medium 49) when a program is executed, data is updated, etc. The CPU 41 executes various processes as described above using a program/data read to the memory 42.

The input device 43 is, for example, a keyboard, a mouse, etc., and is used when an operator, etc. sets the above mentioned various set values (m higher order values, n higher order values, etc.).

The output device 44 can be a display, etc., and displays a setting screen for use by an operator setting the above mentioned set values, a character string recognition result, etc.

The external storage device 45 can be, for example, a magnetic disk device, an optical disk device, a magneto-optical disk device, etc., and stores a program/data, etc. for realization of various functions of the above mentioned character recognition apparatus.

The medium drive device 46 reads the program/data, etc. stored in the portable storage medium 49. The portable storage medium 49 can be,

for example, a FD (floppy disk), CD-ROM, a DVD, magneto-optical disk, etc.

The network connection device 47 is connected to a network, and transmits and receives a program/data to and from an external information processing device.

FIG. 10 shows an example of a storage medium.

As shown in FIG. 10, the program/data can be stored in the portable storage medium 49, loaded into the information processing device, stored in the memory 42, and is then executed. The program/data can be downloaded from a storage device 52 of an external information provider device 51 through a network (Internet, etc.) connected through the network connection device 47.

The present invention can be configured as a storage medium (the portable storage medium 49, etc.) storing the program/data, as a program, as a network (transmission medium) for transmitting the program/data, and as a transmission signal transmitted through the transmission medium when the program/data is downloaded.

In the above mentioned present invention, the following effects can be obtained.

(1) A code string of a key word which is a

node of a character string is automatically extracted from a character string category to be recognized and expressed as a character code. A key word extracted by a key character code extraction unit or a part of it is extracted from a character string image. The character strings in partial areas determined by the extracted key words are holistically recognized. The results of the holistic recognition are verified. Thus, a key word can be automatically extracted from character strings, and the key word can be extracted with high precision although the it is formed by a plurality of characters. Furthermore, since a verifying process is performed on word recognition results, misrecognizing a word as a completely different word can be reduced.

(2) When only a part of a character forming a key word is extracted, the extraction condition as a key character for the preceding and subsequent characters is mitigated, and a character is re-extracted, thereby extracting a key word with high precision although a part of the characters in the key word is considerable deformed.

(3) When the leading and trailing characters in the characters of a key word, and more than a

predetermined ratio of the characters forming the key word are extracted, the partial character string is regarded as a key word, thereby extracting a key word with high precision although
5 characters of the key word touch each other.

(4) When two or more separate characters are extracted in the characters forming a key word, and when more than a predetermined ratio of the characters in an area enclosed by the extracted
10 characters, the partial character string is extracted as a partial character string of the key word, thereby extracting a key word with high precision although the end character touches one of the surrounding characters.

(5) A holistic recognition is performed on an
15 extracted key word or a partial key word, and the probability as a word is verified, thereby extracting a key word with high precision although the character recognition although the character
20 recognition precision is low.

(6) An area segmented as one character is compared in character feature and word feature, and a character string forming part of a key word or the key word is extracted, thereby extracting a key
25 word with high precision although characters in the

key word containing only a small number of characters touch each other.

(7) The recognition precision in word recognition is enhanced by referring to the dictionary in which a word easily misrecognized as a key word is entered as a similar word, thereby extracting a key word with high precision.

(8) When a code string of a key word is extracted, a character having a large number of occurrences in the entire character strings to be recognized, a character having a large number of occurrences in a character string unit, and/or a set of closely associated characters are extracted as key words, thereby automatically extracting a key word from a character string (code) to be recognized.

(9) A character which is not easily misrecognized is entered in advance, and is extracted as a key character when a code string of a key word as a node of a character string is extracted from a character string category, thereby more correctly extracting a key word.

(10) When a word area delimited by a key character/key word is holistically recognized, a word recognizing process is performed, a character

is segmented for the area, and the character is recognized so that a word recognition result can be determined when a character contained in the word recognition result is contained as a higher order
5 and has a number of occurrences equal to or larger than a threshold in the character recognition result, thereby reducing apparent misrecognition.

(11) When a word area is holistically
10 recognized based on a word feature generated by combining character features, the division position of each character in a word image is computed from the matching template, the line density or the peripheral distribution of a word image obtained at each division position is compared with the line
15 density or the peripheral distribution held by each character of a recognized word, and a word recognition result is rejected when a sum of the line density, peripheral distribution, or the difference in the collation ratio is larger than a
20 threshold, thereby reducing the misrecognition of word recognition.

(12) When a word area is holistically
25 recognized based on a word feature generated by combining character features, the number of characters in a recognized word is compared with

the number of characters estimated from a word image, and a word recognition result is rejected when the difference in the number of characters is larger than a threshold, thereby reducing the misrecognition of word recognition.

5

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
222